Benford's Law Analysis on AQI Data of Indian Cities

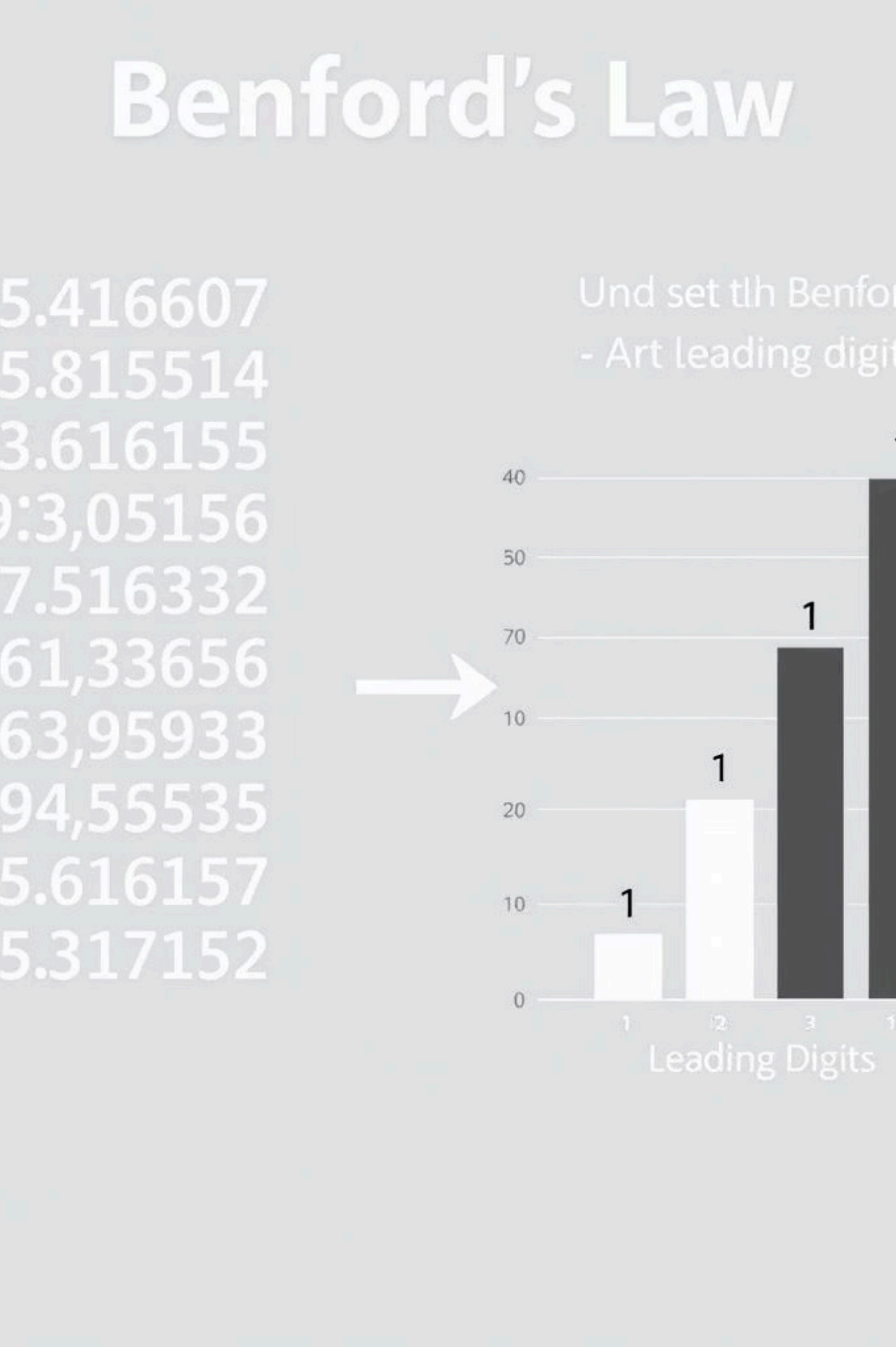
Fundamentals of AI project

Submitted by: Sirex(A N Pavan Sai, Patel Parthkumar, Vidhit T S, Patel Harikrushn)

Date: 4th May 2025

What is Benford's Law?

- Benford's Law is a mathematical principle that predicts the frequency distribution of leading digits in naturally occurring datasets. According to this law, smaller digits occur as the first digit more often than larger ones.
- Key Points:
  a. The number 1 appears as the first digit about 30.1% of the time. Higher digits like 9 appear much less frequently, around 4.6%. It applies to datasets like population numbers, financial data, air quality, and more. Formula:
  b. $P(d) = \log_{10}(1 + 1/d)$ where $P(d)$ is the probability of digit d (1-9) appearing as the first digit.
  c. Why It Matters: Benford's Law is useful for:
     - Identifying natural data patterns Detecting anomalies or possible fraud Verifying authenticity in large datasets



About the Dataset

📄 **Dataset Overview:**

- **Total Records:** 3,168
- **Total Columns:** 11 (including location data, pollutant measurements, and timestamps)

🌍 **Geographic Coverage:**

- Covers **255 unique cities**
- Spans across **532 unique states**
- Data collected from **7 monitoring stations** across India

💨 **Pollutant Information:**

- Tracks multiple pollutants such as **NO_2, SO_2**, and others
- Contains **min, max, and average pollutant values**
- Includes **471 unique pollutant measurements**

📍 **Location Data:**

- Records **latitude and longitude coordinates** for each monitoring station
- Each station mapped to specific geographic locations

✅ **Data Quality:**

- All columns contain **3,168 non-null values** after handling missing data

📊 Attributes Used

📌 **Focused Attribute:**

- **AQI values collected across various Indian cities**

📄 **Dataset Attributes (11 Columns):**

📍 **Location Attributes:**

- **country:** Single value (India)
- **state:** Geographic state location
- **city:** City name
- **station:** Monitoring station name
- **latitude:** Geographic coordinate
- **longitude:** Geographic coordinate

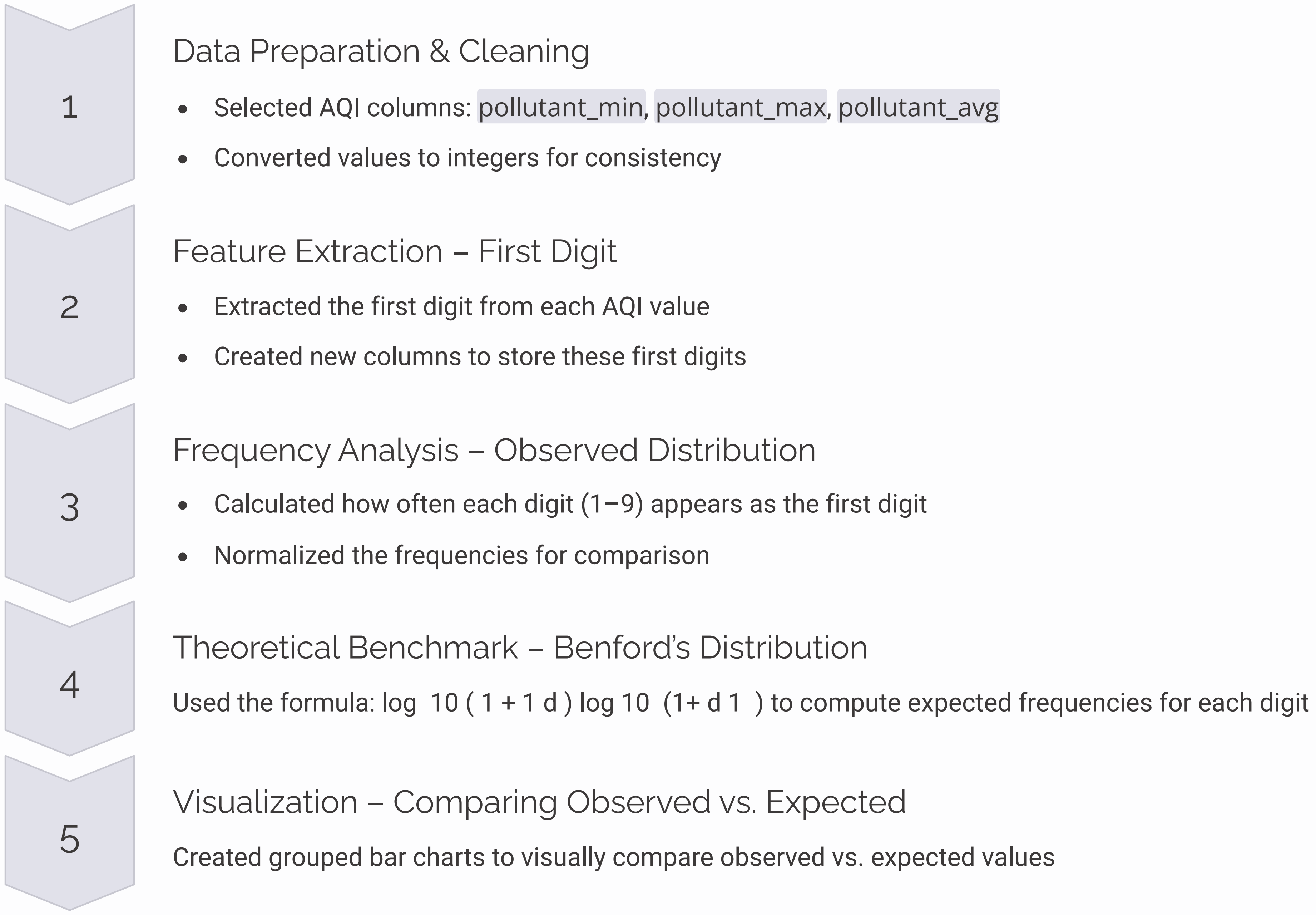
🕑 **Time Attribute:**

- **last_update:** Timestamp of measurements

💨 **Pollutant Measurements:**

- **pollutant_id:** Type of pollutant measured
- **pollutant_min:** Minimum reading
- **pollutant_max:** Maximum reading
- **pollutant_avg:** Average reading

Methodology



1

Data Preparation & Cleaning

- Selected AQI columns: pollutant_min, pollutant_max, pollutant_avg
- Converted values to integers for consistency

2

Feature Extraction – First Digit

- Extracted the first digit from each AQI value
- Created new columns to store these first digits

3

Frequency Analysis – Observed Distribution

- Calculated how often each digit (1–9) appears as the first digit
- Normalized the frequencies for comparison

4

Theoretical Benchmark – Benford's Distribution

Used the formula: log 10 (1 + 1 d) log 10 (1+ d 1) to compute expected frequencies for each digit

5

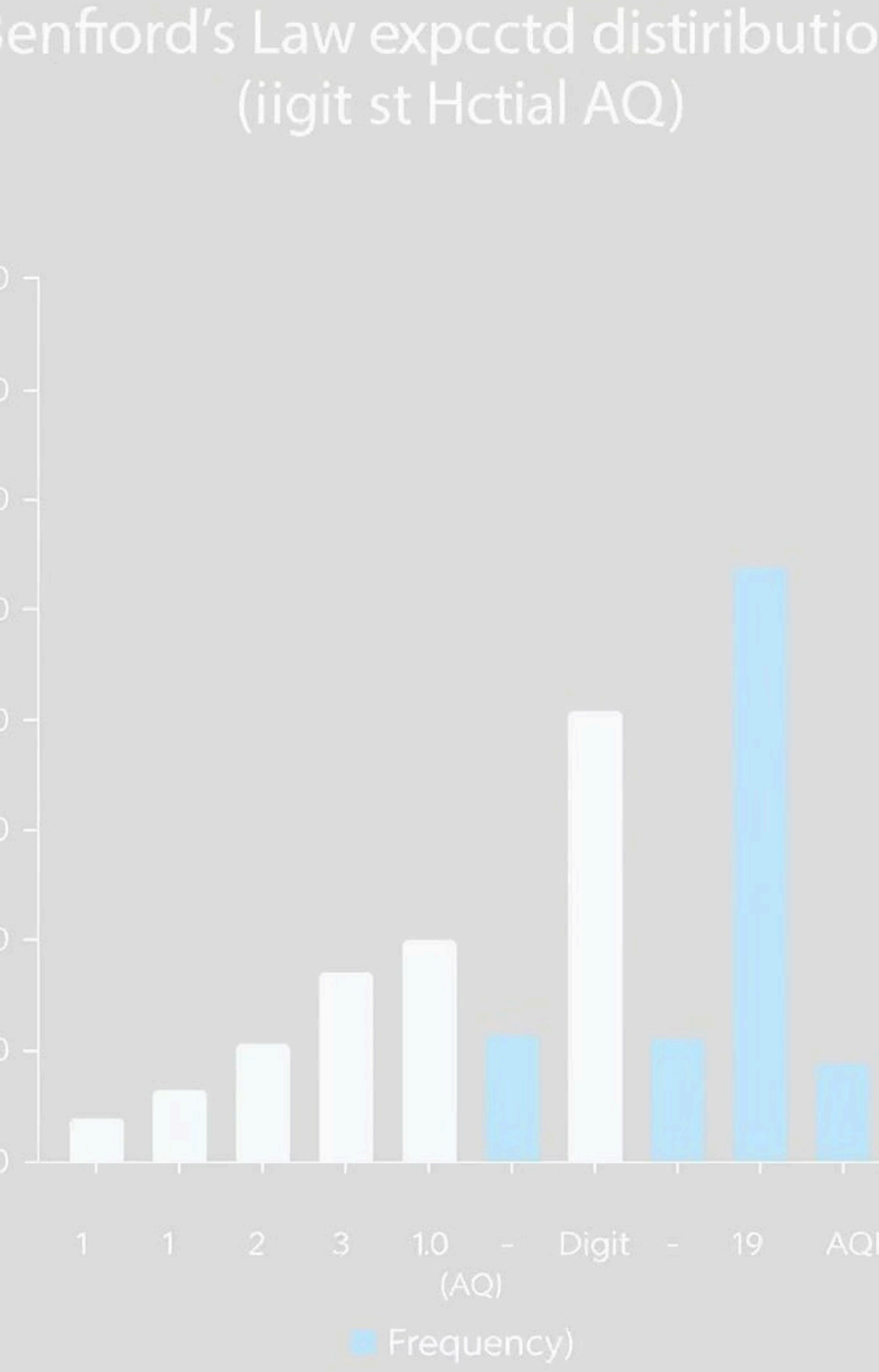
Visualization – Comparing Observed vs. Expected

Created grouped bar charts to visually compare observed vs. expected values

Tools: Python with NumPy,Pandas, Plotly

Expected vs Actual Distribution

Digit	Expected %	Actual %
1	30.1%	24.25%
2	17.6%	19.58%
3	12.49%	14.64%
4	9.69%	11.27%
5	7.92%	8.98%
6	6.69%	7.65%
7	5.80%	5.93%
8	5.12%	4.38%
9	4.58%	3.32%



Insights & Interpretation

📊 Overall Fit

🔍 Fit Assessment:

- **Moderate deviation overall** — data isn’t a perfect fit but follows recognizable, systematic patterns

📉 Key Deviations:

- **Digit 1:** Largest under-representation (**-5.85%**)
- **Middle digits (2–6):** Consistently over-represented
- **Digit 7:** Nearly perfect fit (**+0.13%**)
- **Digits 8 & 9:** Slight under-representation

📌 Overall Summary:

- **Moderate fit** to Benford’s Law
- **Systematic deviations** suggest structured data collection
- Some natural alignment, especially with **digit 7**

📊 Deviations & Observations

⚠️ Moderate Deviations:

- Data shows **systematic deviations** from Benford’s Law
- **First Digit (1):**
    - **Under-represented by -5.85%**
    - **Expected:** 30.10%, **Actual:** 24.25%
    - *Possible cause:* Measurement thresholds or rounding practices
- **Middle Digits (2–6):**
    - Consistently **over-represented**
    - Largest in **digit 3** (+2.15%)
    - *Suggests clustering in measurements*

✅ Minor Deviations:

- **Digit 7:**
    - Closest to expected value
    - **Deviation:** +0.13%
    - *Indicates natural occurrence at this level*

Insights & Interpretation

📊 Possible Causes

🏙️ Data Collection Methods:

- Measurement equipment calibration ranges
- Rounding or truncation practices
- Standardized measurement protocols

🌿 Environmental Factors:

- Pollution level reporting thresholds
- Natural limits in pollutant concentrations
- Regulatory compliance targets

⚙️ Technical Reasons:

- Instrument precision limits
- Systematic measurement intervals
- Data processing and cleaning methods

📌 Other Factors:

- Urban pollution clustering
- Seasonal variations affecting pollutant levels

📊 Significance & Insights

📍 Dataset Coverage:

- **3,168 records** across **255 cities** and **532 states**
- Data from **7 monitoring stations**, up to **2025**

📊 Data Quality:

- Complete, well-structured data with **11 attributes**
- Consistent formats and accurate geographic coverage

💨 Measurement Patterns:

- Tracks multiple pollutants (**NO_2, SO_2**, etc.)
- Standardized min, max, avg readings at regular intervals

📈 Benford's Analysis Insights:

- Systematic **deviation in digit 1** (24.25% vs 30.10%)
- **Middle digits over-represented**
- **Digit 7 naturally aligned**

🌏 Environmental & Analytical Value:

- Supports **policy decisions** and **air quality monitoring**
- Enables **trend analysis**, **geographic** and **temporal insights**

📊 Conclusion

- **AQI data shows a moderate fit with Benford’s Law**, with systematic deviations
- Highlights **consistent measurement patterns** likely due to **reporting practices, environmental thresholds, and instrument constraints**
- Confirms **hidden trends in air quality data distribution behavior**
- Demonstrates potential for **using Benford’s analysis as a diagnostic tool** for **data quality assessment and anomaly detection** in environmental monitoring

Individual contributions

1

A N Pavan Sai

Data Preparation & Cleaning

- Understand the dataset structure
- Handle missing values / NaNs
- Convert data types as required
- Ensure data is in clean, numeric format for analysis

Delivered: Cleaned, structured dataset ready for feature extraction

2

Patel Parthkumar

Feature Extraction & Frequency Analysis

- Create a function to extract the first digit from each numeric value
- Calculate frequency of each first digit (1–9)
- Normalize frequencies using relative proportions

Delivered: Observed frequency distribution of first digits

3

Patel HariKrushn

Benford’s Law Calculation & Visualization

- Compute theoretical Benford’s Law distribution (log10(1 + 1/d))
- Compare observed vs expected frequencies
- Create grouped bar charts (e.g., with Plotly/Matplotlib)
- Highlight deviations

Delivered: Visuals comparing AQI data with Benford’s expected pattern

4

Vidhit T S

EDA, Insights, Interpretation & Conclusion

- Perform additional EDA:
    - Summary statistics of AQI data (min, max, avg values)
    - City/state-wise AQI distribution
    - Most/least polluted cities
- Identify key deviations and patterns from Benford’s analysis
- Create final PPT slides and compile the report/notebook

Delivered: EDA visualizations, insights summary, conclusion, and final presentation

Thank You*